Knihovna pro určení vzájemně podobných fotografií vhodného pro produkční provoz

Bc. Dobroslav Pelc

Diplomová práce
2018

Univerzita Tomáše Bati ve Zlíně
Fakulta aplikované informatiky

*** Nascanované zadání, strana 1 ***

*** Nascanované zadání, strana 2 ***

Prohlašuji, že

- beru na vědomí, že odevzdáním diplomové práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb. o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších právních předpisů, bez ohledu na výsledek obhajoby;
- beru na vědomí, že diplomové práce bude uložena v elektronické podobě v univerzitním informačním systému dostupná k prezenčnímu nahlédnutí, že jeden výtisk diplomové práce bude uložen v příruční knihovně Fakulty aplikované informatiky. Univerzity Tomáše Bati ve Zlíně a jeden výtisk bude uložen u vedoucího práce;
- byl/a jsem seznámen/a s tím, že na moji diplomovou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
- beru na vědomí, že podle § 60 odst. 1 autorského zákona má UTB ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
- beru na vědomí, že podle § 60 odst. 2 a 3 autorského zákona mohu užít své dílo – diplomovou práci nebo poskytnout licenci k jejímu využití jen připouští-li tak licenční smlouva uzavřená mezi mnou a Univerzitou Tomáše Bati ve Zlíně s tím, že vyrovnání případného přiměřeného příspěvku na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše) bude rovněž předmětem této licenční smlouvy;
- beru na vědomí, že pokud bylo k vypracování diplomové práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tedy pouze k nekomerčnímu využití), nelze výsledky diplomové práce využít ke komerčním účelům;
- beru na vědomí, že pokud je výstupem diplomové práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá. Neodevzdání této součásti může být důvodem k neobhájení práce.

Prohlašuji,

- že jsem na diplomové práci pracoval samostatně a použitou literaturu jsem citoval. V případě publikace výsledků budu uveden jako spoluautor.
- že odevzdaná verze diplomové práce a verze elektronická nahraná do IS/STAG jsou totožné.

Ve Zlíně .....

podpis autora

ABSTRAKT

Cílem této práce je analýza možností pro určení vzájemně podobných fotografií. Na základě analýzy student vybere nejvhodnější návrh řešení pro potřeby reálného produkčního provozu. Výslednou komponentu realizuje formou distribuované služby.

Klíčová slova: Přehled klíčových slov

ABSTRACT

Text of the abstract

Keywords: Some keywords

Zde je místo pro případné poděkování, motto, úryvky knih, básní atp.

OBSAH

ÚVOD ... **9**

I TEORETICKÁ ČÁST ... **9**

1 POUŽITÉ ANALYTICKÉ METODY ... **11**

1.1 Diferenční analýza ... 11

2 KLASIFIKACE ŘEŠENÝCH OBLASTÍ ... **12**

2.1 Konstruktivní změny fotografie ... 12

2.1.1 Návrh řešení ... 12

2.1.2 Oblast zájmu ... 12

2.2 Destruktivní změny fotografie ... 12

2.2.1 Návrh řešení ... 12

2.2.2 Oblast zájmu ... 12

2.3 Kombinace konstruktivních a destruktivních změn ... 13

2.3.1 Návrh řešení ... 13

2.3.2 Oblast zájmu ... 13

2.4 Výběr reprezentativního vzorku ... 13

2.4.1 Návrh řešení ... 13

2.4.2 Oblast zájmu ... 13

3 DALŠÍ NADPIS ... **14**

3.1 Podnadpis ... 14

3.1.1 Podpodnadpis ... 14

3.1.2 Podpodnadpis ... 14

II ANALYTICKÁ ČÁST ... **14**

4 BRAINSTORMING ... **16**

4.1 Akademický kruh ... 16

4.2 Profesní kruh ... 16

5 DIFERENČNÍ ANALÝZA (GAP ANALÝZA) ... **17**

5.1 Popis současného stavu ... 17

5.2 Popis cílového stavu ... 17

5.2.1 Nefunkční požadavky ... 17

5.3 Rozdíly ... 17

5.4 Návrh variant k dosažení cíle ... 18

5.4.1 Výpočet koeficientů na CPU ... 18

5.4.2 Výpočet koeficientů na CPU s paralelizací na GPU ... 18

5.4.3 Výpočet koeficientů na PC farmě ... 18
5.5 Zhodnocení variant ... 18
5.5.1 Výpočet koeficientů na CPU ... 18
5.5.2 Výpočet koeficientů na CPU s paralelizací na GPU ... 19
5.5.3 Výpočet koeficientů na PC farmě ... 19
6 BENCHMARKING ... 20
6.1 Testovací prostředí ... 20
6.1.1 Použitý HW ... 20
6.1.2 Použitý SW ... 20
6.2 Výsledky testování ... 20
6.2.1 Výpočet koeficientů na CPU ... 21
6.2.2 Výpočet koeficientů na CPU s paralelizací na GPU ... 21
6.2.3 Výpočet koeficientů na PC farmě ... 22
III PROJEKTOVÁ ČÁST ... 23
7 DISTRIBUOVANÁ SLUŽBA ... 25
7.1 Fronta nezpracovaných obrázků ... 25
7.2 Servlet pro stažení obrázků ... 25
7.3 Odeslání výsledků ... 25
8 KLIENT ... 26
ZÁVĚR ... 27
SEZNAM POUŽITÉ LITERATURY ... 28
SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK ... 29
SEZNAM OBRÁZKŮ ... 30
SEZNAM TABULEK ... 31
SEZNAM PŘÍLOH ... 32

ÚVOD

První odstavec pod nadpisem se neodsazuje, ostatní ano (pouze první řádek, odsazení vertikální mezy odstavci je typycké pro anglickou sazbu; czech babel toto respektuje, netřeba do textu přidávat jakékoliv explicitní formátování, viz ukázka sazby tohoto textu s následujícím odstavcem).

Formátování druhého odstavce. Text text text text text text text text text text text text.

I. TEORETICKÁ ČÁST

1 Použité analytické metody

1.1 Diferenční analýza

Diferenční analýzu (někdy též Gap analýza) navrhl Igor Ansoff. Skládá se z následujících kroků:

- Popis stávajícího stavu
- Stanovení cílů (popis cílového stavu)
- Určení rozdílu (mezery) mezi stávajícím a cílovým stavem
- Návrh variant dosažení cílového stavu (alternativní strategie)
- Zhodnocení variant a výběr nevhodnější z nich
- V případě potřeby se celý postup opakuje, dokud není dosaženo cílového stavu

2 Klasifikace řešených oblastí

TODO: navodit téma + uvézt základní škálu

2.1 Konstruktivní změny fotografie

TODO: definovat konstruktivní změny

2.1.1 Návrh řešení

Křížová korelace zrychlená pomocí diskrétní Fourierovy transformace. TODO: rozvést

2.1.2 Oblast zájmu

TODO: uvézt

Změny vlastností

- Změna barev; TODO: Popis změny barev
- Změna kontrastu;TODO: Popis změny kontrastu
- Změna jasu;TODO: Popis změny jasu

Změny obsahu

- Vodotisk
- Logo
- Šum

2.2 Destruktivní změny fotografie

TODO: definovat destruktivní změny

2.2.1 Návrh řešení

Výpočet Hausdorfovy vzdálenosti mezi konvexními polyedry, které reprezentující hrany v obrazu. TODO: rozvést

2.2.2 Oblast zájmu

- Změna komprese (rozmazaná fotka)
- Změny rozlišení
  - Ořez (v jedné nebo obou dimenzích
  - Deformace (v jedné nebo obou dimenzích)

2.3 Kombinace konstruktivních a destruktivních změn

TODO: Obecně je porovnání problematické, rozvést.

2.3.1 Návrh řešení

Redukce fotografie na její prahovou velikost jako příprava na křížovou korelaci viz konstruktivní změny.

2.3.2 Oblast zájmu

- Asymetrická změna obou stran s čímkoliv
- Změna kvality v důsledku zhoršeni komprese s čímkoliv
- Logo nebo vodotisk v kombinaci s předcházejícími

2.4 Výběr reprezentativního vzorku

Pro skupinu vzájemně si podobných fotografií vybereme nejvhodnějšího kandidáta, který bude následně ostatní fotografie zastupovat.

2.4.1 Návrh řešení

Pro tyto účely zavedeme koeficient podobnosti, který zjednodušeně určíme jako VELIKOST × OSTROST + JAS. Výsledek bude na intervalu <0, 1>. Hodnoty blížící se nule mají nejnižší koeficient zastupitelnosti a naopak hodnoty blížící se 1 mají nejvyšší koeficient zastupitelnosti.

2.4.2 Oblast zájmu

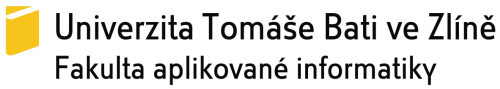
Střední hodnota jasu Pouze zohlednění zda fotka není příliš jasná nebo příliš tmavá, jednoduchý algoritmus

Poměrný počet hran Test "rozmazanosti" fotky, konvoluce s vhodným (nutné testování) jádrem typu horní propust (s celkovým součtem 0) - tedy výsledek neutrální podklad z kterého "vystupují" hrany, velké množství hran ⇔ fotka není rozmazaná

Rozlišení fotografie Pouze klasická velikost fotky (větší ⇔ lepší).

3 Další nadpis

Tato sekce obsahuje ukázku vložení obrázku (Obr. 3.1).



Obr. 3.1 Popisek obrázku

3.1 Podnadpis

Tato sekce obsahuje ukázku vložení tabulky (Tab. 3.1).

Tab. 3.1 Popisek tabulky

	1	2	3	4	5	Cena [Kč]
F	(jedna)	(dva)	(tři)	(čtyři)	(pět)	300

3.1.1 Podpodnadpis

3.1.2 Podpodnadpis

Citace knihy. [1]

II. ANALYTICKÁ ČÁST

4 Brainstorming

Analytický metoda sloužící ke sběru myšlenek, námětů a případné zevrubné konstruktivní kritice dané problematiky. V rámci této práce byla použita v akademickém a profesním kruhu pro identifikaci základních ukazatelů pro další kroky analýzy.

4.1 Akademický kruh

Diskutovány zejména technické možnosti. Kde a jak lze vůbec porovnání fotografií provádět strojově. K další analýze vyly vyb

4.2 Profesní kruh

V kruhu s provozovatelem byly kladeny nejvyšší nároky na flexibilitu a propustnost celého řešení.

5 Diferenční analýza (GAP analýza)

5.1 Popis současného stavu

Je požadováno porovnání rastrových bitmap za účelem identifikace vzájemně podobných fotografií. Řešení je hledáno pro produkční provoz. Konzumentem cílového řešení je webový portál dovolena.cz, který má přibližně dva miliony fotografií. Průměrný počet přístupů k některé fotografii je přibližně 100 přístupů za sekundu. Jedná se především o hotely a jejich okolí. Webový portál slouží spíše jako datový konsolidátor. Nabídka portálu značně ovlivňuje cílové portfolio fotografií. Podle testovacích měření se za jeden týden obmění cca 10% fotografií z celkového množství. Jako testovací vzorek byl vybrán jeden nejmenovaný hotel a jeho 139 fotografií. Zákazník webového portálu vidí všechny fotografie v nesetříděné galerii. Některé fotografie jsou unikátní, ale většina si je velmi podobná. U některý dokonce nejsou lidským okem patrné rozdíly.

5.2 Popis cílového stavu

Konsolidované fotografie prezentované klientovi budou v maximální možné míře obsahovat unikátní fotografie. Vzájemně si podobné fotografie budou odfiltrovány a zůstane pouze jedna a to fotografie s nejvyšším indexem zastupitelnosti. Klient nebude čekat na zpracování podobnosti obrázků. Buď budou zpracované, pak klient uvidí jen unikáty, nebo ne a pak uvidí vše v původním stavu. V takovém případě se poměrově zvýší priorita na výpočet podobnosti fotek tohoto hotelu vůči ostatním ve frontě na výpočet. Cílové řešení musí být schopno operovat řádově s jednotkami milionů fotografií s týdenní fluktuací 15%.

5.2.1 Nefunkční požadavky

- Bezúdržbový systém
- Nevyžadující v průběhu času další financování
- Minimální vstupní investice
- Maximální kompatibility s aktuálním HW

5.3 Rozdíly

- Nově vznikne nástroj pro určení koeficientu podobnosti dvou fotografií.
- Nově vznikne nástroj pro určení koeficientu zastupitelnosti vzájemně si podobných fotografií.
- Fotografie jednoho hotelu budou oindexovány a vnitřně škálovány do skupin pomocí koeficientů výše.
- Dojde k navýšení celkového počtu fotografií.
- Zvýší se také fluktuace fotografií (na očekávaných 15%).

5.4 Návrh variant k dosažení cíle

Pilířem celého řešení bude backendová strana cílového konzumenta. Limity a také jednotlivé možnosti pro realizaci jsou velmi omezeny nutností integrovat do současného řešení. Nejen z těchto důvodů má serverová strana spíše podpůrný charakter v projektu jako celku. Její význam je spíše v propojení všech jednotlivých komponent. Dojde tedy k modifikaci existujícího produkčního server-side prostředí. Nově zde poběží služba, která bude

- poskytovat zadání na určení koeficientů podobnosti a zastupitelnosti,
- poskytovat metadata nezbytná pro distribuci výpočtu,
- konzumovat výsledek distribuované operace,
- kompletovat zpracovaná data do cache vhodné pro silný organický provoz.

Naopak klientská strana je naprosto autonomní. Pro realizaci lze použít jak libovolnou platformu, tak libovolné technologie. Jediným technickým limitem je schopnost standardizovaným způsobem komunikovat se serverovou stranou.

5.4.1 Výpočet koeficientů na CPU

Základní myšlenka je využít nejdostupnější produkční HW a na zavedeném serveru spustit novou službu. Hlavní výhodou je dostupnost produkčního HW ve vlastním datacentru cílového konzumenta. Podstatnou nevýhodou fakt, že pro určení koeficientů výše není CPU ideální platforma.

5.4.2 Výpočet koeficientů na CPU s paralelizací na GPU

Základní myšlenka je osadit do produkčního serveru pracovní grafickou kartu a počítat podobnost fotografií na GPU.

5.4.3 Výpočet koeficientů na PC farmě

Základní myšlenka je využít HW osobních PC, kterých je v každé větší firmě požehnaně a neprovádět výpočet na jednom stroji, na každém dostupném stroji.

5.5 Zhodnocení variant

Zhodnocení vychází z benchmarkingu, který byl vyhodnocen jako pomocná analýza.

5.5.1 Výpočet koeficientů na CPU

- Za necelých 6 dní spočítá týdenní přírůstek.
- Za zbylou dobu z týdenního cyklu dále vypočítá přibližně 3,7% z celkového objemu sto milionu koeficientů.

- Pro plné vypočítání všech koeficientů potřebuje dalších cca 27 týdnů.
- Rezerva pro další růst (navýšení celkového počtu fotografií) je cca 18%.

5.5.2 Výpočet koeficientů na CPU s paralelizací na GPU

- Za necelých 8 hodin spočítá týdenní přírůstek.
- Za část ze zbylé doby týdenního cyklu dále vypočítá 100% z celkového objemu sto milionu koeficientů.
- Pro plné vypočítání všech koeficientů potřebuje celkem 2 dny.
- Rezerva pro další růst (navýšení celkového počtu fotografií) je cca 300%.

5.5.3 Výpočet koeficientů na PC farmě

Výsledek jednoho kancelářského PC je zanedbatelný a nemůže se rovnat předchozím variantám. Vezmeme-li v úvahu, že těchto strojů je k dispozici 16 hodin denně více než 300 kusů, začíná to již vypadat v číslech jinak.

- Za 2,5 hodiny spočítá týdenní přírůstek (za předpokladu 300 aktivních PC).
- Za část ze zbylé doby týdenního cyklu dále vypočítá 100% z celkového objemu sto milionu koeficientů.
- Pro plné vypočítání všech koeficientů potřebuje celkem 16 hodin, což je v tomto případě 1 den.
- Rezerva pro další růst (navýšení celkového počtu fotografií) je cca 700%.

Samozřejmě s tímto nestandardním krokem je spojená také revize infrastruktury, především kvalita a šířka pásma sítě, centrální správa PC aj, které je nutné v tuto chvíli zanedbat.

6 Benchmarking

Jednotlivé kandidáty pro realizaci klientské části (rozvedené v GAP analýze) podrobíme výkonnostním a srovnávacím testům. Základní předpoklady:

- Řádově je celkem potřeba jednorázově odbavit 100 000 000 výpočtů obou koeficientů.
- Na týdenní bázi následně odbavovat denní přírůstky (řádově 15 000 000 výpočtů obou koeficientů).
- Na denní bázi je to tedy něco přes 2 000 000 výpočtů obou koeficientů.

6.1 Testovací prostředí

6.1.1 Použitý HW

Simulace produkčního serveru Při volbě HW k testování byla zvolena pracovní stanice Lenovo, která je svoji sestavou velmi blízko produkčnímu serveru. Vzhledem k tomu, že disponuje starší generací CPU, bylo do ní osazeno i odpovídají GPU, aby bylo možné výsledky aproximovat na reální produkční HW.

- Lenovo ThinkStation D20
- Operační systém Windows 10 pro 64 bit
- CPU Intel(R) Xeon(R) X5670 @ 2.93 GHz (2×cpu, 6×core, 12×thread)
- RAM 32 GB DDR3 (1066)
- GPU NVIDIA GeForce GTX 460
- HDD SAS 10.000 ot

Simulace klasického PC Alternativní varianta řešení je využít v rozumné míře osobní PC. Cílový konzument řešení disponuje několika více než 300 ks kancelářských PC. Jako testovací vzorek byl vybrán zástupce drtivé většiny těchto PC.

- Dell optiplex 780
- Operační systém Windows 7 pro 32 bit
- CPU Intel Core 2 Duo E7500 2,93 GHz
- RAM 4 GB DDR3
- HDD SATA 250 GB

6.1.2 Použitý SW

V úvodní fázi projektu probíhal vývoj i testování algoritmů v Mathlabu.

6.2 Výsledky testování

Byly realizovány dva testovací scénáře.

- Výpočet čistě na CPU s maximálním využitím HW.
- Výpočet na CPU s využitím GPU pro paralelizaci vybraných procesů.

6.2.1 Výpočet koeficientů na CPU

Výpočet na CPU byl spuštěn v pěti vláknech. Z počátku vypadal průběh velmi slibně. Bohužel při delším testu se výpočet začal značně propadat. Později se ukázalo, že hlavním důvodem je odložené uvolňování RAM. Jinými slovy dokud byla další volná RAM, výpočet jel velmi rychle. S potřebou uvolnit RAM se výpočet řádově snížil.

- Krátkodobý test
  - doba: 5 minut
  - počet opakování testu: 10
  - průměrně za sekundu: 205 výpočtů obou koeficientů
- Dlouhodobý test
  - doba: 6 hodin
  - počet opakování testu: 10
  - průměrně za sekundu: 31 výpočtů obou koeficientů
- Propad o 85%

Na obrázku níže (Obr. 6.1) je výstup z profilování výpočtu na CPU. Obsahuje poměrově zvýrazněné dlouho trvající operace vůči zbytku procesu. V rámci profilování bylo zpracováno deset tisíc iterací a doby jednotlivých částí zprůměrovány.

Obr. 6.1 Vizualizace procesu výpočtu koeficientů na CPU

Celková zátěž alokovaného HW nebyla příliš slavná. Nepodařilo se efektivně využít CPU, které dosahovalo průměrné zátěže 30% zejména kvůli neustálé realokaci operační paměti, na kterou čekalo zkrátka vše. Úvodní myšlenka využít hrubou sílu produkčního HW se ukázala jako velmi špatná.

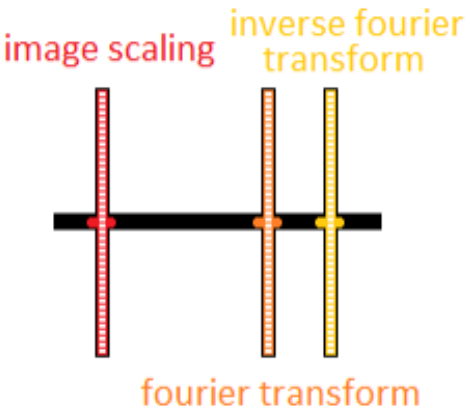
6.2.2 Výpočet koeficientů na CPU s paralelizací na GPU

Výpočet byl spuštěn v jednom vlákně na CPU. Vhodné operace pro GPU jsou delegovány pro paralelní zpracování. Nejprve bylo na GPU paralelizován pouze úvodní scaling obrázků. To nemělo příliš valný dopad na výsledky. Následně byly paralelizovány na GPU také prováděné transformace obrázků. Právě tento krok měl zásadní vliv na zrychlení celé operace. Je zde opět patrný lehký propad krátkodobého testu oproti dlouhodobému.

- Krátkodobý test
  - doba: 5 minut

  - počet opakování testu: 10
  - průměrně za sekundu: 617 výpočtů obou koeficientů
- Dlouhodobý test
  - doba: 6 hodin
  - počet opakování testu: 10
  - průměrně za sekundu: 559 výpočtů obou koeficientů
- Propad o 10%

Na obrázku níže (Obr. 6.2) je výstup z profilování výpočtu na CPU s využitím GPU pro paralelizací dlouhotrvajících operací na CPU. Obsahuje poměrově zvýrazněné dlouho trvající operace vůči zbytku procesu. V rámci profilování bylo zpracováno deset tisíc iterací a doby jednotlivých částí zprůměrovány.



Obr. 6.2 Vizualizace procesu výpočtu koeficientů na CPU s paralelizací na GPU

Zátěž na CPU dosahuje v průměru 5%. Naopak GPU dosahuje zátěže cca 75%. Otázkou zůstává, jak dlouho je schopná GPU pracovat pod tímto permanentním zatížením.

6.2.3 Výpočet koeficientů na PC farmě

Výpočet byl spuštěn v jednom pracovním vlákně s omezenou možností alokace RAM na 1GB. Jedná se v podstatě o alternativu výpočtu obou koeficientů čistě na CPU. Stejně tak tomu odpovídal i celkový průběh zpracování, který téměř stejný, pouze s nižší dotací vypočítaných dvojic koeficientů. Lze tedy přejít rovnou na výsledky, jelikož samotné zpracování nic nového nepřineslo.

- Krátkodobý test
  - doba: 5 minut

  - počet opakování testu: 10
  - průměrně za sekundu: 10 výpočtů obou koeficientů
- Dlouhodobý test
  - doba: 6 hodin
  - počet opakování testu: 10
  - průměrně za sekundu: 6 výpočtů obou koeficientů
- Propad o 40%

Zátěž na CPU je konstantní. Jedno jádro ze dvou jede permanentně na 100%. Celková zátěž CPU tedy 50%. Vzhledem k tomu, že stoj již netrpěl na realokaci paměti, je v případě PC úzkým hrdlem opravdu CPU. Ostatní sledované metriky (vytížení HDD, síťový provoz, pracovní teplota) se příliš nevychýlili z normálu.

III. PROJEKTOVÁ ČÁST

7 Distribuovaná služba

7.1 Fronta nezpracovaných obrázků

7.2 Servlet pro stažení obrázků

7.3 Odeslání výsledků

8 Klient

ZÁVĚR

Text závěru

SEZNAM POUŽITÉ LITERATURY

[1] *Český chmel: atlas odrůd* [online]. [cit. 2004-10-15]. Dostupný z WWW: `http://www.beer.cz/humulus/`.

SEZNAM POUŽITÝCH SYMBOLŮ A ZKRATEK

ABC	Význam zkratky

SEZNAM OBRÁZKŮ

Obr. 3.1 Popisek obrázku 14
Obr. 6.1 Vizualizace procesu výpočtu koeficientů na CPU 21
Obr. 6.2 Vizualizace procesu výpočtu koeficientů na CPU s paralelizací na GPU 22

SEZNAM TABULEK

Tab. 3.1 Popisek tabulky 14

SEZNAM PŘÍLOH

P I. Název přílohy

PŘÍLOHA P I. NÁZEV PŘÍLOHY

Obsah přílohy